

Device–Circuit–System Co-Design of Reconfigurable Oxide-Semiconductor 2T Non-Volatile Gain-Cell Memory and Sandwich-3D LLM Accelerator

Luqi Zheng, Chang Niu, Shuting Du, Sumi Lee, Amir Massah Bavani, Zehao Lin, Peide D. Ye, and Haitong Li

¹Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

This work was supported by the U.S. National Science Foundation FuSe2 program (Award No. 2425498).

ABSTRACT In this work, we present an oxide-semiconductor (OS)-based 2T non-volatile gain-cell (NVGC) memory and a “sandwich 3D” stacking scheme, co-designed using ALD InOx OSFETs and OS-FeFETs, for efficient large language model (LLM) acceleration. The new 2T-NVGC enables native, in-memory reconfiguration of static/dynamic kernels for Transformers and context-switching workloads, while realizing efficient FP/INT compute-in-memory. With full-stack, cross-layer co-designs, we present the first complete mixed-signal physical design implementation of an OS-on-CMOS 3D-IC with 32.9 TOPS/W energy efficiency and 102.4 TOPS/mm² compute density. Post-3D-layout simulations calibrated with device experiments demonstrate that the 3D-NVGC LLM accelerator delivers a 52.3× gain in holistic system efficiency over planar CMOS-based and a 3.2× gain over 3D DRAM-based solutions, respectively.

INDEX TERMS Oxide Semiconductor, FeFET, Non-Volatile Gain-Cell Memory, Compute-in-Memory, Floating-Point Acceleration, Reconfigurable Memory, 3D-IC, LLM Accelerator, Device-Circuit Co-Design, OS-on-CMOS Integration

I. INTRODUCTION

LARGE language models (LLMs) driving remarkable advances across diverse areas of computing and intelligence demand efficient hardware solutions without sacrificing precision. State-of-the-arts that exploit promising “architectural recipes” such as compute-in-memory (CIM) and 3D [1] [2] [3] [4] [5], and “technological ingredients” including back-end-of-line (BEOL) logic (e.g., amorphous oxide semiconductor transistors) and memories (e.g., FeRAM/FeFETs) [5-6], still face several major challenges, as summarized in Fig.1: (Challenge 1) Transformer’s attention mechanism requires both static and dynamic matrix operations with evolving model/task-dependency. Such fundamental fine-grained reconfigurability is not yet attainable with any incumbent or emerging memories; (Challenge 2) Floating-point (FP) formats, crucial for reliable LLM inference and training, lack efficient in-memory realization even in advanced CIM designs due to weight pre-alignment padding overheads [1]

[2]; and (Challenge 3) 3D memory architectures with 3D V-NAND/Fe-NAND [3], [4] and 3D DRAM [5] restrict all logic and peripheral circuits to the CMOS-under-array (CuA) tier, ultimately limiting compute efficiency due to confined silicon footprint with intensive, intermediate data movement down to CuA.

To tackle all three challenges simultaneously, we adopt a device–circuit–system co-design approach experimentally harnessing ALD InOx-based OSFETs and OS-FeFETs. A new 2T-NVGC memory is developed to enable reconfigurable static/dynamic kernels natively in memory for context-switching LLM adaptability, while realizing high-utilization FP data storage and high-efficiency FP computation. Co-optimizing NVGC memory and InOx mixed-signal circuits, we further develop an OS-Memory/OS-Logic/OS-Memory “sandwich 3D” architecture for scalable BEOL-integrated AI acceleration. Unlike conventional CuA 3D memory architectures, where peripheral and logic circuits

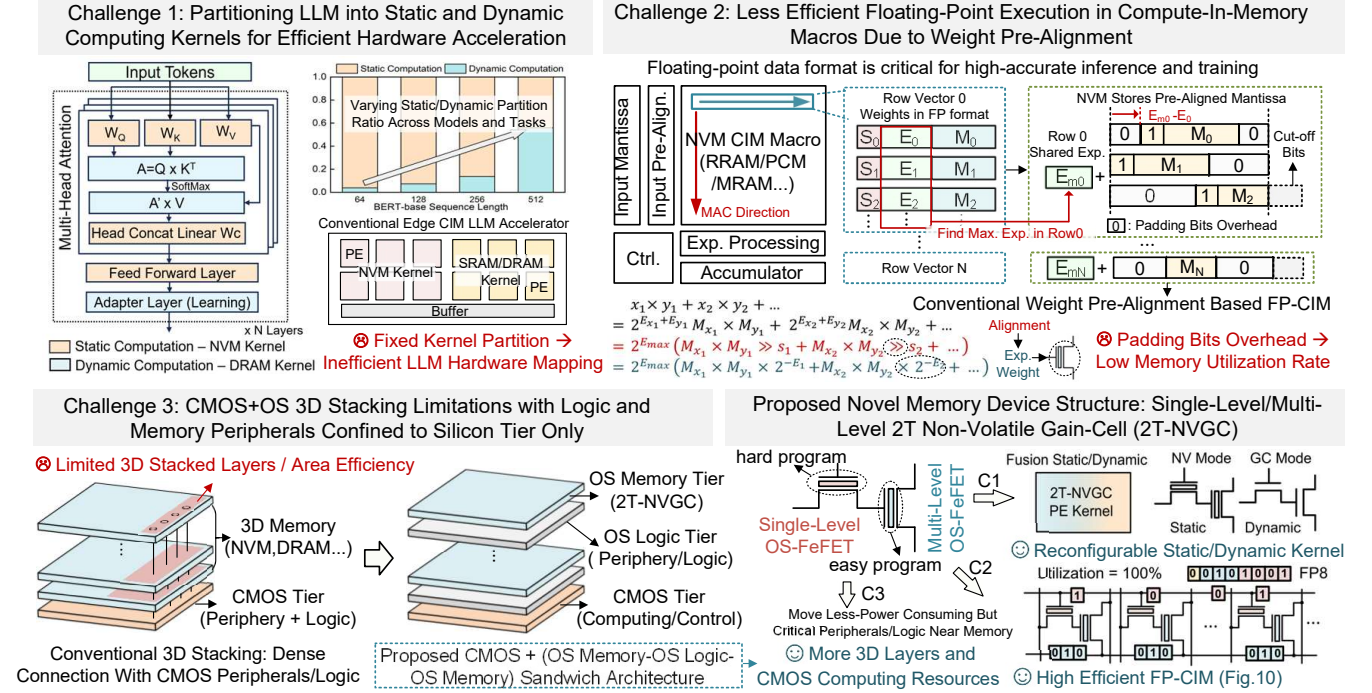


FIGURE 1. Key challenges in LLM accelerators, and our full-stack co-design solution harnessing OS technologies.

remain confined to a single silicon CMOS tier, the proposed “sandwich 3D” paradigm vertically distributes memory arrays, mixed-signal peripherals, and lightweight logic across multiple oxide-semiconductor BEOL tiers. Such an architecture is uniquely enabled by oxide-semiconductor technologies, which simultaneously support both memory devices (OS-FeFETs/OS-eDRAM) and logic transistors (InOx OS-FETs) within BEOL-compatible fabrication processes. This organization fundamentally differs from prior 3D memory stacking approaches by enabling each vertically integrated memory tier to become computation-capable, rather than functioning solely as passive storage. As a result, the proposed architecture alleviates CuA routing congestion and intermediate data movement bottlenecks while improving compute scalability, memory utilization, and inter-tier communication efficiency for future large-scale LLM accelerators. With our experimentally-calibrated OS standard cell library and OS process design kit (PDK), this work provides the first complete mixed-signal physical design implementation of a 3D OS-on-CMOS LLM accelerator, considering practical implications of high-performance NMOS-only OSFETs and a relaxed OS node on silicon.

II. NON-VOLATILE GAIN-CELL MEMORY

Leveraging the low-leakage characteristics of wide-bandgap oxide semiconductors, OSFETs are well-suited for 2T gain-cell eDRAMs with separate access and read transistors [8] [9]. Compared with conventional silicon-based gain-cell memories, oxide-semiconductor devices provide improved retention characteristics while maintaining BEOL compat-

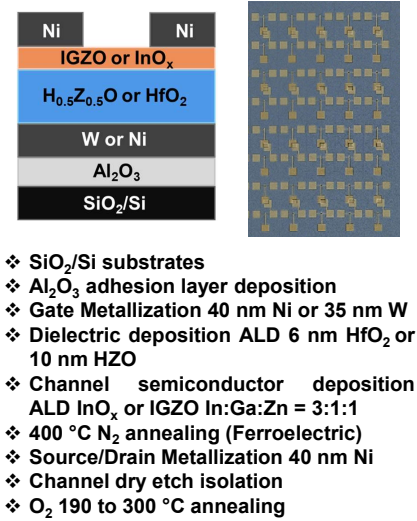


FIGURE 2. Device schematic, image, and fabrication flow of ALD InOx-based FETs and FeFETs.

ibility for monolithic 3D integration. Our 2T-NVGC retains this compact topology but replaces both transistors with OS-channel ferroelectric FETs (OS-FeFETs), enabling native reconfiguration between dynamic (charge-based) and non-volatile modes for efficient LLM hardware mapping and context-switching workloads. Compared with a 1T-1FeFET cell [10], the 2T-NVGC offers significant circuit-level advantages when addressing the device-level challenges in programming the access-FeFET in the 2-FeFET GC topology,

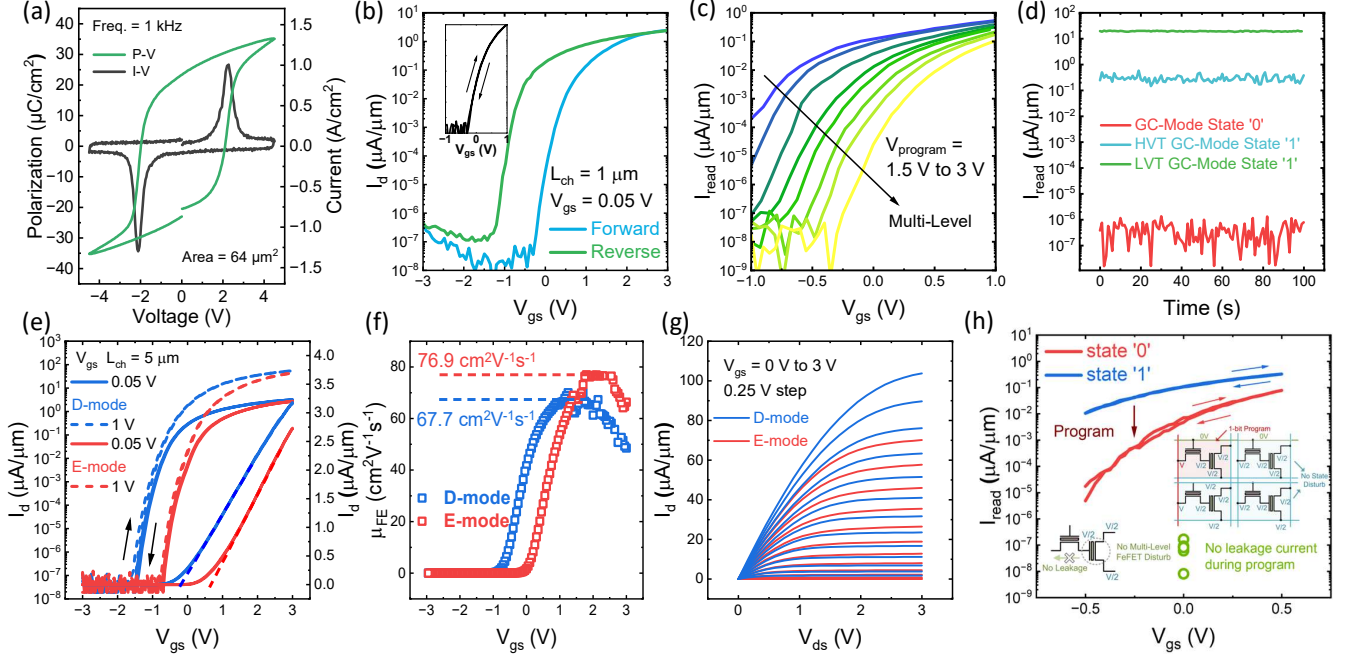


FIGURE 3. (a) Polarization vs. voltage characterization of ALD IGZO Fe-Cap. (b) IGZO FeFET transfer curves. Inset: no hysteresis below 1 V V_{gs} . (c) Multi-level characterization of OS-FeFET with 8 distinct states. (d) Gain-cell DRAM mode charge-based retention characterization. (e) E-mode and D-mode InOx FETs transfer characteristics. (f) E-mode and D-mode InOx FETs field-effect mobility characteristics. (g) E-mode and D-mode InOx FETs output characteristics. (h) Programming single-level access FeFET without leakage or disturb. The measurement results are used for circuit simulations.

while also improving flexibility for mixed FP/INT CIM operations.

To leverage 2T-NVGC for FP-CIM, as shown in Fig. 1, exponent bits are stored in the read-FeFETs, enabled by reliable multi-bit storage via read-verify programming, while mantissa bits are stored bit-wise in the access FeFETs, optimized for single-bit one-shot programming. This structure realizes exponent-weighted mantissa FP-CIM, where exponents encoded in the read-FeFET states are directly reflected in the read current, eliminating the need for mantissa pre-alignment and padding overhead in state-of-the-art CIM circuits. Since exponent differences beyond 8 can be truncated with negligible accuracy loss, a 3-bit read-FeFET with 8 states is sufficient to encode most FP formats used in practical LLM acceleration. The NVGC devices and OS logic transistors are fabricated and characterized to support subsequent 3D accelerator circuit design.

The device structure, optical image and fabrication flow of the ALD InOx-based logic and memory devices (InO₃-FETs and IGZO-FeFETs) are illustrated in Fig. 2. The gate electrode consists of either a 40nm Ni or a 35nm W layer. The gate dielectric is formed by plasma-enhanced ALD (PE-ALD) of 6nm HfO₂ or 10nm Hf_{0.5}Zr_{0.5}O₂ (HZO), deposited at 200°C. The semiconducting IGZO channel is deposited by thermal ALD at 225°C, followed by the formation of 40nm Ni source/drain contacts. For OS-FeFETs, a 400°C annealing step in N₂ is applied after film deposition to enable the ferroelectricity. Channel isolation is achieved

through dry etching, and post-metal annealing is conducted at temperatures ranging from 190°C to 300°C.

Fig. 3(a) presents the polarization–voltage characteristics of the ferroelectric–semiconductor–metal capacitor, indicating a high-quality ferroelectric gate stack. Fig. 3(b) shows the standard transfer characteristics of an ALD IGZO-FeFET, demonstrating distinct memory states and a large memory window. The inset highlights the absence of hysteresis under ± 1 V bias, confirming a non-destructive read operation. Fig. 3(c) illustrates the multi-level programming capability of the NVGC read-FeFET using programming voltages from 1.5 V to 3 V, yielding eight clearly distinguishable states. Fig. 3(d) presents the retention characteristics of the GC mode of NVGC, with the access-FeFET in both low-V_{th} (LVT) and high-V_{th} (HVT) states, exhibiting stable charge retention without degradation up to 100 seconds. Logic circuits based on ALD InOx can be constructed using enhancement-mode (E-mode) and depletion-mode (D-mode) FETs. Fig. 3(e)–(g) shows the transfer and output characteristics of devices with a 5 μ m channel length. The E-mode and D-mode FETs exhibit high mobilities of 76.9 cm²/V·s and 67.7 cm²/V·s, respectively. Fig. 3(h) shows successful single-bit programming of the NVGC access-FeFET without leakage or disturbance to the read-FeFET or adjacent cells. The inset illustrates array-level programmability using a modified V/2 scheme adapted from the 1T-1FeFET design [10], enabling reliable reverse programming from LVT to HVT.

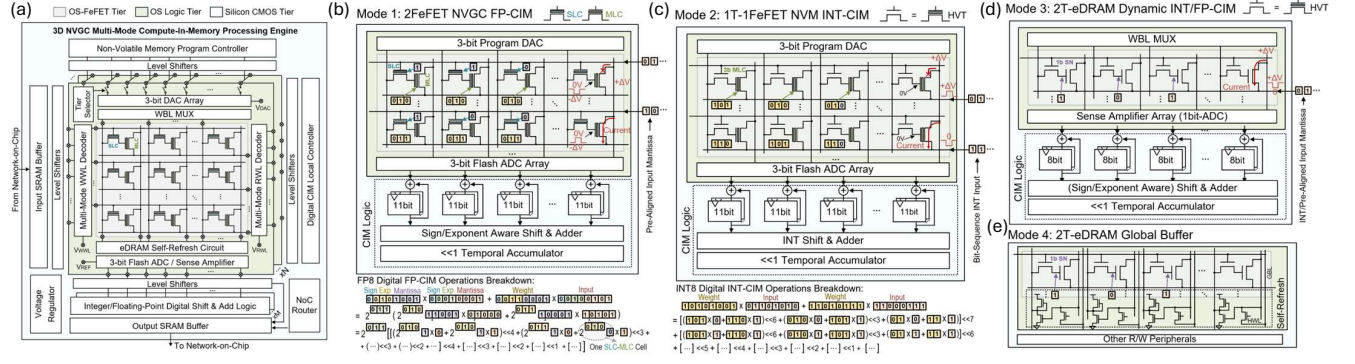


FIGURE 4. (a) Floorplan and circuit design of 3D NVGC CIM PE across OS-FeFET, OS logic and Si CMOS tiers. Operation modes of the NVGC PE: (b) Mode 1, NV-FP mode: 2T-NVGC configured as a fused single-level multi-level 2-FeFET cell for FP CIM. (c) Mode 2, NV-INT mode: 2T-NVGC configured as 1T-1FeFET cell for INT CIM. (d) Mode 3, GC mode: 2T-NVGC configured as 2T-GC eDRAM cell for INT/FP CIM. (e) Mode 4, Buffer mode: self-refreshing eDRAM global buffer.

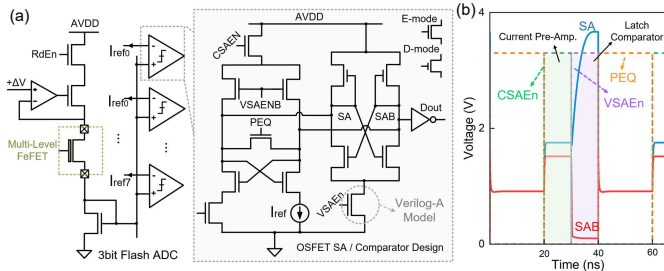


FIGURE 5. Analog circuit design with InOx transistors. (a) NMOS-only 3-bit flash ADC incorporating E-mode and D-mode OSFETs, with eight comparators that also function as sense amplifiers. (b) Simulated waveform of a two-stage current sense amplifier using a custom Verilog-A model with measured characteristics of fabricated OSFETs.

III. 3D LLM ACCELERATOR CIRCUIT DESIGN

Utilizing the proposed NVGC memory on a heterogeneous CMOS+OS platform in a monolithic 3D integration scheme, we perform complete analog, digital, and mixed-signal 3D-IC implementation with full place-and-route (PnR) across the CMOS and interwoven OS tiers. Unlike many prior BEOL-based accelerator demonstrations that mainly focus on device characterization or isolated circuit blocks, this work emphasizes practical physical implementation considerations including tier partitioning, mixed-voltage integration, routing congestion, inter-tier communication, and realistic area overheads. Such physical-design-level considerations are particularly important for evaluating the actual scalability and efficiency of BEOL-augmented 3D AI accelerators.

First, the overall processing engine (PE) architecture is illustrated in Fig.4(a). Each PE spans multiple vertically integrated tiers including the NVGC memory tier, the OS logic tier, and the Si CMOS compute tier. To balance performance, power, and routing complexity, only essential low-power peripheral circuits such as row decoders, self-refresh controllers, and local sensing interfaces are implemented on the OS logic tier. This partitioning minimizes long-distance routing between the OS and CMOS domains while reducing the overall power overhead associated with peripheral oper-

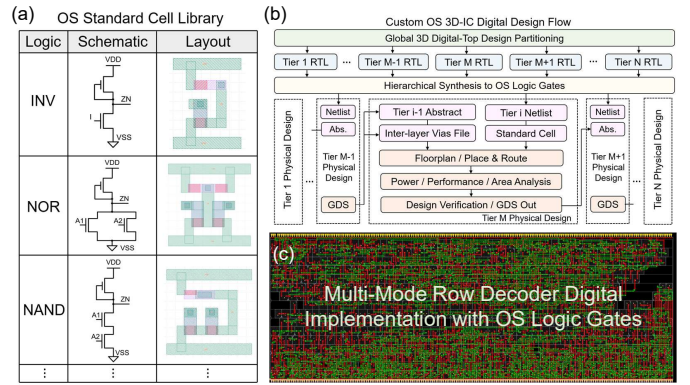


FIGURE 6. Digital circuit design with custom OS PDK. (a) Standard cell library based on InOx transistors. (b) Customized 3D-IC digital design flow. (c) Example result: multi-mode row decoder with OS gates.

ations. Analog front-end circuits including ADCs and DACs for near-memory processing are shared across multiple bit-lines through analog multiplexing structures to further improve area and energy efficiency. The main digital compute engines, scheduling logic, and control units remain on the CMOS tier to leverage the superior device performance and routing capability of advanced CMOS technology. Since the OS logic operates at a higher voltage domain, dedicated level shifters are incorporated to interface the low-voltage CMOS logic with the 3.3 V OS circuits.

Next, we illustrate the operation modes and dataflows of the NVGC PE in Fig.4(b)–(e). The proposed architecture supports four dynamically reconfigurable operating modes, including static FP/INT CIM using NV-FeFET storage, dynamic FP/INT CIM using GC-eDRAM operation, and an eDRAM buffer mode for storing intermediate activations and temporary data during inference and training. This reconfigurable operation allows the same physical memory fabric to efficiently support workloads with different temporal characteristics and precision requirements. Static NV-mode operation is particularly suitable for storing long-term model weights and reusable parameters, while dynamic GC-mode

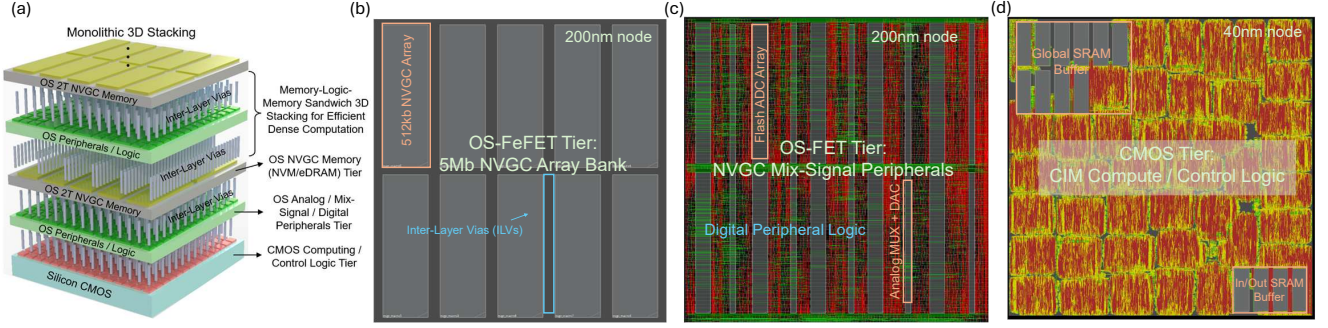


FIGURE 7. (a) Illustration of 3D NVGC accelerator featuring OS memory/OS logic/OS memory sandwich tiers on silicon CMOS. (b) OS-FeFET tier: 5 Mb NVGC array bank, composed of 512 Kb sub-arrays, connected to the OS logic tier through inter-layer vias. (c) OS-FET tier: analog components (such as DACs, flash ADCs) with digital peripherals and essential functional logic. (d) CMOS tier: CIM auxiliary and control logic for upper tiers, system scheduler, global buffer and input/output SRAM buffer.

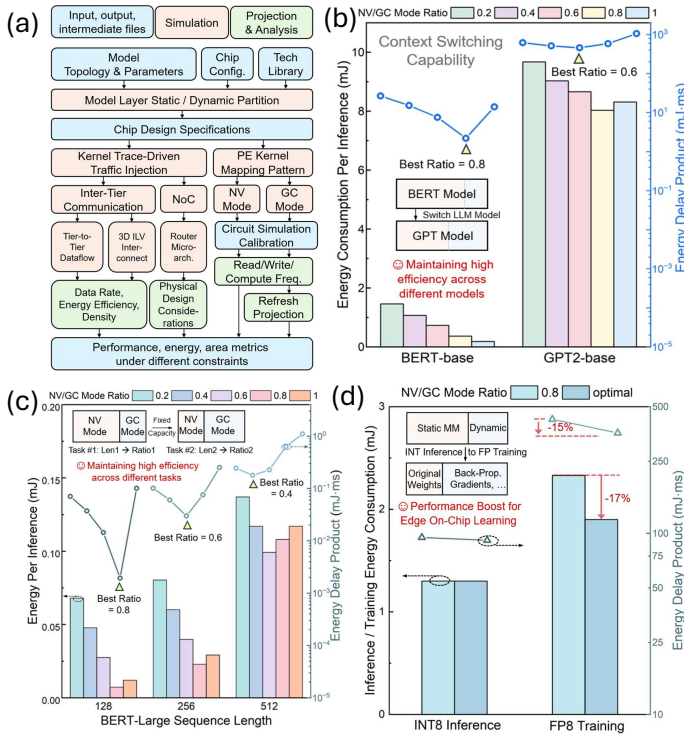


FIGURE 8. (a) Full-stack system DSE and benchmarking framework. (b) BERT evaluation on 3D NVGC with adaptive NV/GC mode partitioning, revealing task-specific EDP-optimal configurations. (c) Efficient on-demand context switching across BERT/GPT-2 models via adapting NV/GC ratios, guided by system inference energy and EDP. (d) Reconfigurable NVGC PE reduces FP8 training energy by 17% for BERT, enabled by INT/FP adaptation with minimal overhead.

operation enables efficient handling of activations, attention matrices, and temporary training states. Such flexibility is especially important for Transformer-based workloads where the ratio between static and dynamic data can vary significantly across models and sequence lengths.

Specifically, as shown in Fig.4(b) capturing the ‘NV-FP mode’, the 2T-NVGC is configured as a fused single-level/multi-level (SLC/MLC) 2-FeFET cell to support floating-point CIM. In this mode, multi-level read-FeFET

states encode exponent information while access-FeFETs store mantissa bits, enabling exponent-aware analog current accumulation without explicit mantissa pre-alignment. The accumulated outputs are digitized by the shared flash ADC array and processed by sign/exponent-aware shift-and-add logic together with temporal accumulation units. Fig.4(c) shows the ‘NV-INT mode’, where the same 2T-NVGC structure is reconfigured into a 1T-1FeFET style INT-CIM array for efficient integer inference with simplified bit-serial accumulation. Fig.4(d) presents the ‘GC mode’ for dynamic INT/FP-CIM operations, where the 2T-NVGC functions as a charge-based eDRAM compute fabric capable of handling dynamic activations, attention matrices, and temporary training states in the charge domain efficiently. In this mode, the flash ADC array is reconfigured into single-bit sense-amplifier arrays together with exponent-aware digital accumulation logic to support both INT and FP operations. Since dynamic weights and activations stored in the GC mode are frequently updated and rewritten, conventional weight pre-alignment introduces relatively small overhead compared with stationary NV-mode weights, making GC mode more suitable for flexible dynamic computation. Finally, Fig.4(e) illustrates the Buffer mode, where the 2T structure operates as a self-refreshing eDRAM global buffer with lightweight refresh circuitry for intermediate data storage and local communication across PE stages.

At the architectural level, these four modes allow the PE to dynamically adapt the same physical memory array to distinct computational roles, including long-term weight storage, high-utilization FP-CIM, conventional INT-CIM, dynamic activation processing, and temporary buffering. Such multi-modal operation substantially improves memory utilization and efficiencies compared with fixed-function CIM architectures, while reducing both the demand for large on-chip SRAM buffers and the volume of inter-tier data movement.

To validate circuit functionality across multiple abstraction levels, all PE operations are verified at register-transfer-level (RTL) through mixed-signal simulations using experi-

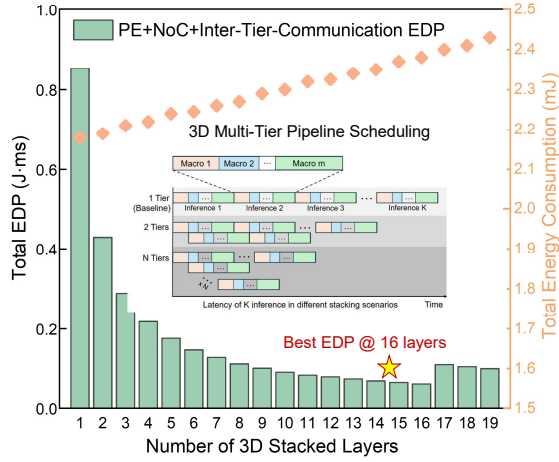


FIGURE 9. 3D stacking design exploration considering CMOS resources, NoC, and inter-tier communication with cross-tier pipelining, showing best performance with 16 layers.

mentally calibrated behavioral models of the NVGC arrays. These simulations capture both digital control behavior and analog memory characteristics including multi-level FeFET programming states, gain-cell retention behavior, and analog sensing operations. Fig.5(a) presents the analog implementation of a 3-bit flash ADC constructed using E-mode and D-mode OSFETs. The comparator employs a two-stage architecture consisting of a current-mode pre-amplifier followed by a latch-based sense amplifier to improve sensing speed and robustness under OS device variations. Measured device I-V characteristics are used to extract Verilog-A models for accurate circuit-level simulation. The simulated waveform shown in Fig.5(b) demonstrates high-speed rail-to-rail sensing with minimal offset and stable operation across multiple sensing cycles.

Fig.6(a) further presents the custom OS standard cell library developed for digital synthesis and place-and-route on the OS logic tier. Since conventional CMOS libraries are not directly compatible with NMOS-only oxide-semiconductor logic, the library that we developed is specifically optimized for ALD InOx devices and incorporates experimentally calibrated transistor-level characteristics. Cell performance, power, and area (PPA) metrics are extracted through detailed transistor-level simulation and integrated into the digital design flow. Fig.6(b)(c) illustrate the overall 3D-IC design methodology and a representative multi-mode row decoder synthesized using the proposed OS logic library, demonstrating the feasibility of large-scale digital implementation on the OS tier.

Leveraging the proposed “sandwich 3D” stacking scheme for the 3D NVGC accelerator (Fig.7(a)), our OS-on-CMOS physical implementation consists of three tightly integrated functional tiers that collaboratively distribute memory, mixed-signal processing, and control functionality across the vertical dimension. Unlike conventional CuA-based 3D memory architectures, where vertically stacked layers mainly

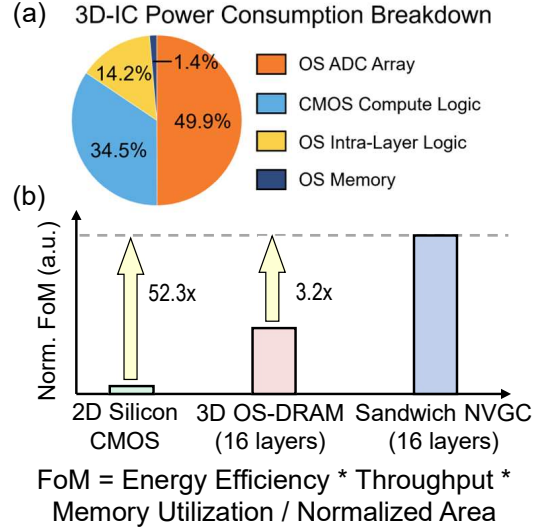


FIGURE 10. (a) 3D-IC cross-tier power breakdown. (b) FoM comparison among 2D silicon CMOS, 3D OS-DRAM, and this work.

increase storage capacity while computation and peripherals remain confined to the silicon tier, the proposed architecture allows each OS tier to participate directly in local data processing and orchestration. Such an organization is enabled by the unique capability of oxide-semiconductor technologies to simultaneously support BEOL-compatible logic and memory devices within the same integration platform.

The top FeFET memory tier contains NVGC memory banks connected to the OS logic tier through dense inter-layer vias (Fig.7(b)), enabling high-bandwidth local communication between storage and mixed-signal processing circuits. The intermediate OS logic tier integrates mixed-signal peripheral circuits including row decoders, local controllers, sensing interfaces, ADCs/DACs, and lightweight digital logic (Fig.7(c)), thereby offloading substantial local dataflow operations from the bottom CMOS tier. The bottom CMOS tier implements the main compute engines, scheduling logic, NoC infrastructure, and global control circuitry (Fig.7(d)), leveraging the superior performance and routing capability of advanced CMOS technology for large-scale digital orchestration.

This vertical partitioning transforms the stacked memory hierarchy from passive storage layers into computation-capable tiers with localized processing capability. As a result, the proposed architecture simultaneously improves compute density, routing scalability, and memory utilization efficiency while reducing long-distance data movement and CuA congestion bottlenecks. Such insights suggest a scalable direction for future BEOL-integrated AI systems, where memory, mixed-signal peripherals, and lightweight logic are co-distributed across vertically stacked tiers rather than centralized in a single silicon layer.

The OS tiers are implemented using a relaxed 200 nm technology node while the CMOS tier utilizes a commercial

TABLE 1. Comparison with prior 2D/3D AI hardware designs using OS-eDRAM or NVM. This work achieves ultra-high compute density and supports LLM context/task switching via inherent reconfigurability, enabling both inference and training across integer and floating-point formats.

Reference	IEDM'23 [8]	ISSCC'23 [12]	IEDM'24 [9]	ISSCC'24 [13]	VLSI'24 [14]	This Work
Cell Type	2T0C	3T1C	2T0C	1T1R	STT-MRAM	2T NVGC
Multi-Bits / Memory	1b / OS-eDRAM	1b / OS-eDRAM	1b / OS-eDRAM	1b / RRAM	1b / MRAM	3b / OS-FeFET
Memory Retention	Volatile	Volatile	Volatile	Non-Volatile	Non-Volatile	Non-Volatile / Dynamic Reconfig.
Data Format	INT	INT	INT	INT	INT/FP	INT/FP
ML Workload	CNN	Endpoint AI	Transformer	Robotics	CNN	LLM
Inference/Training	Inference	Inference	Inference	Inference	Inference	Inference/Training
3D Memory Bank	Yes	Yes	Yes	No	No	Yes
BEOL Peripherals	No	Yes	No	No	No	Yes
3D-Layer Logic	No	No	No	No	No	Yes
System-Technology Co-Optiomized	No	No	No	No	No	Yes
Compute Density (TOPS/mm ²)	50	NR ^{*1}	0.007	0.013	0.018	102.4
Energy Efficiency ^{*2} (TOPS/W)	256	4.44	0.099	1.14 ^{*3}	63.9	32.9

^{*1} NR = Not reported. ^{*2} Normalized to 8-bit operations. ^{*3} RRAM storage + SRAM compute-in-memory.

40 nm foundry process, enabling realistic evaluation of lab-to-fab scaling implications for future BEOL-integrated AI accelerators. The complete design is implemented at the processing engine (PE) level, with inter-PE communication handled through a network-on-chip (NoC) infrastructure. For modeling simplicity, intra-PE communication across vertically integrated OS tiers is abstracted as NoC-level communication in the system-level evaluation framework. This abstraction allows scalable evaluation of large accelerator configurations while preserving the key effects of inter-tier communication overhead and vertical integration constraints.

IV. SYSTEM-LEVEL FEATURES AND INSIGHTS

To faithfully benchmark LLM inference and continual learning workloads at the system level, we develop a multi-tier 3D modeling and design space exploration (DSE) framework (Fig.8(a)), adapted from our recent work [11]. Unlike simplified analytical estimations that only consider isolated compute kernels, our framework captures interactions across multiple abstraction levels including LLM model characteristics, accelerator architecture, circuit implementation, and device technology parameters. The DSE tool supports operation partitioning across different memory and compute resources, models PE-level kernel mapping and dataflow scheduling with inter-/intra-tier communication overheads, and projects end-to-end power, performance, and area (PPA) metrics under realistic workload conditions. This framework further enables comprehensive system-technology co-optimization (STCO) analysis for the proposed 3D LLM accelerator exploiting the reconfigurable NVGC memory and the “sandwich 3D” architecture.

First, the reconfigurable NVGC memory enables on-demand optimization of energy and energy-delay-product

(EDP) across LLM workloads with varying sequence lengths (Fig.8(b)). Unlike prior designs with fixed memory compositions, the proposed architecture dynamically adjusts the ratio between non-volatile (NV) and gain-cell (GC) operating modes according to workload behavior, thereby eliminating unnecessary FeFET rewrite overheads and OSFET refresh overheads. Longer sequence lengths tend to favor lower NV/GC ratios because they introduce a larger proportion of dynamic matrix operations associated with attention computation and activation updates. In contrast, shorter sequence lengths benefit from higher NV ratios due to increased reuse of static model parameters. This workload-adaptive reconfiguration capability enables the accelerator to maintain near-optimal efficiency across a broad range of LLM operating conditions.

Second, the accelerator supports rapid context and task switching through dynamic reconfiguration of the NV/GC memory composition. As shown in Fig.8(c), the optimal NV/GC ratio for minimizing energy consumption and EDP shifts significantly when the workload changes from BERT to GPT-2, reflecting fundamental differences in model architecture, dataflow characteristics, and temporal memory behavior. Such adaptability is particularly important for emerging edge and continual-learning applications, where heterogeneous AI tasks may execute sequentially on the same hardware platform. The proposed NVGC architecture therefore provides a flexible hardware substrate capable of efficiently supporting diverse Transformer-based models without requiring dedicated accelerator redesign.

Third, the proposed system simultaneously supports energy-efficient INT-based inference and FP-based training workloads through the inherent flexibility of the NVGC memory architecture. As illustrated in Fig.8(d), FP train-

ing workloads require a larger proportion of GC-mode memories to accommodate dynamic weight updates during backpropagation and gradient accumulation, while inference-oriented workloads benefit from larger NV-mode allocation for static weight storage. By enabling seamless transitions between these operating modes, the accelerator efficiently supports both inference and on-device training within a unified hardware framework, which is increasingly important for future adaptive and personalized AI systems. Finally, we implement pipelined inference with multi-tier scheduling on the proposed 3D architecture while evaluating vertical scalability from an STCO perspective. The analysis considers tradeoffs among compute density, communication overhead, tier utilization, and CMOS area constraints across the OS-FeFET memory tier, OS logic tier, and CMOS compute tier. As shown in Fig.9, a pareto-optimal operating point achieving the best balance between EDP and compute density is obtained at approximately 16 vertically stacked layers. Beyond this point, additional layers introduce diminishing returns due to increasing communication overhead and limited CMOS resource scalability. With the figure-of-merit (FoM) defined in Fig.10, the proposed 3D NVGC accelerator achieves a $52.3\times$ improvement in holistic system efficiency over planar CMOS-based counterparts and a $3.2\times$ improvement over 3D DRAM-based solutions for LLM hardware acceleration. Table 1 further summarizes comparisons with state-of-the-art emerging memory-based AI accelerators. Compared with prior works primarily targeting fixed-function INT inference with either volatile or non-volatile memories, the proposed 3D NVGC architecture uniquely combines dynamic/non-volatile reconfigurability, native FP/INT CIM capability, BEOL-integrated peripherals and logic, and full-stack system-technology co-optimization within a monolithic 3D implementation. These features enable efficient support for LLM inference, training, and context-switching workloads while simultaneously achieving ultra-high compute density and practical mixed-signal physical design scalability for future 3D AI accelerators.

V. CONCLUSION

We presented a device-circuit-system co-designed 3D LLM accelerator based on oxide-semiconductor logic and non-volatile gain-cell memories. The key contributions of this work are threefold: (1) a new reconfigurable 2T-NVGC memory with native FP compute capability, enabling seamless transitions between dynamic and non-volatile operating modes; (2) a new “sandwich 3D” BEOL stacking scheme for scalable and high-density 3D CIM architectures; and (3) the first full-PnR monolithic 3D LLM accelerator that captures the practical lab-to-fab implications of OS logic and memory technologies. Through device characterizations for experimental calibration, mixed-signal circuit implementation, and full-stack system-level evaluation, the co-designed solution efficiently supports both integer and floating-point LLM workloads across inference and training. Together, this

work establishes a scalable, LLM-native technology platform that can co-evolve and scale with future 3D-ICs.

ACKNOWLEDGMENT

This work was supported by the U.S. National Science Foundation under Award No. 2425498 with industry partners as specified in the Future of Semiconductors (FuSe2) program.

REFERENCES

- [1] T. -H. Wen *et al.*, “A 22nm 16Mb Floating-Point ReRAM Compute-in-Memory Macro with 31.2TFLOPS/W for AI Edge Devices,” *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2024, pp. 580-582.
- [2] L. Zheng *et al.*, “CENTAUR: A 38.5-TFLOPS/W 600MHz Floating-Point Digital Compute-In-Memory Engine with 40nm Fusion RRAM-eDRAM Macros Featuring 3D-MAC Operation,” *2025 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Daejeon, Korea, Republic of, 2025, pp. 139-141.
- [3] Z. Huo, W. Cheng and S. Yang, “Unleash Scaling Potential of 3D NAND with Innovative Xtacking® Architecture,” *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, HI, USA, 2022, pp. 254-255.
- [4] J. -G. Lee *et al.*, “Analog Computation in Ultra-High Density 3D FeNAND for TB-level Hyperscale AI Models,” *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2024, pp. 1-4.
- [5] K. S. Choi *et al.*, “A Three Dimensional DRAM (3D DRAM) Technology for the Next Decades,” *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, HI, USA, 2024, pp. 1-2.
- [6] C. Niu *et al.*, “First Demonstration of BEOL Wafer-Scale All-ALD Channel CFETs Using IGZO and Te for Monolithic 3D Integration,” *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2024, pp. 1-4.
- [7] Z. Lin *et al.*, “High-Performance BEOL-Compatible Atomic-Layer-Deposited In2O3 Fe-FETs Enabled by Channel Length Scaling down to 7 nm: Achieving Performance Enhancement with Large Memory Window of 2.2 V, Long Retention 10 years and High Endurance 108 Cycles,” *2021 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2021, pp. 17.4.1-17.4.4.
- [8] W. Tang *et al.*, “30 Mb/mm²/layer 3D eDRAM Computing-in-Memory with Embedded BEOL Peripherals and Local Layer-wise Calibration based on First-Demonstrated Vertically-stacked CAA-IGZO 4F2 2T0C Cell,” *2023 International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2023, pp. 1-4.
- [9] S. Liu *et al.*, “Edge Continual Training and Inference with RRAM-Gain Cell Memory Integrated on Si CMOS,” *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2024, pp. 1-4.
- [10] S. G. Kirtania *et al.*, “Amorphous Indium Oxide Channel FeFETs with Write Voltage of 0.9V and Endurance 1012 for Refresh-Free 1T-1FeFET Embedded Memory,” *2024 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2024, pp. 1-4.
- [11] S. Du *et al.*, “3D-CIMlet: A Chiplet Co-Design Framework for Heterogeneous In-Memory Acceleration of Edge LLM Inference and Continual Learning,” *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, San Francisco, CA, USA, 2025, pp. 1-7.
- [12] Y. Yakubo *et al.*, “Crystalline Oxide Semiconductor-based 3D Bank Memory System for Endpoint Artificial Intelligence with Multiple Neural Networks Facilitating Context Switching and Power Gating,” *textit 2023 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 12-14.
- [13] S. D. Spetalnick *et al.*, “A 40nm VLIW Edge Accelerator with 5MB of 0.256pJ/b RRAM and a Localization Solver for Bristle Robot Surveillance,” *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2024, pp. 482-484.
- [14] D. -Q. You *et al.*, “A 22nm Nonvolatile AI-Edge Processor with 21.4TFLOPS/W using 47.25Mb Lossless-Compressed-Computing STT-MRAM Near-Memory-Compute Macro,” *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, HI, USA, 2024, pp. 1-2.